Otmane El Bourki

AI Engineer · Production LLM & RAG · Agentic AI · Inference Optimization · Cloud & MLOps · Evaluation

Carrières-sous-Poissy, Île-de-France · France · Authorized to work in France

+33 7 75 73 57 51 • otmane.elbourki@gmail.com • linkedin.com/in/oelbourki • github.com/oelbourki • oelbourki.com

AI Engineer (confirmé) with **3 years of professional and independent experience** building LLM, RAG and agentic systems (**4.5+ years in AI/ML** including research). Specialized in production LLM, RAG and agentic AI, with experience spanning retrieval, model serving, inference optimization, orchestration, cloud deployment and evaluation. Delivered **3×** **LLM inference throughput** at **47% lower cost**, built multi-agent systems for finance, and currently develop AI accelerators within ALTEN's **A3**.

Experience

AI Engineer Confirmé — A^3 (AI at ALTEN) June 2026 – Present
ALTEN Group — AI for Management System & Accelerators Rabat, Morocco

- **Product delivery: live in 3 months, 101 active users.** Built and productionized an AI references-discovery platform (React, FastAPI, pgvector, Azure) for ALTEN's sales teams and directors; 1,103 searches in early rollout.
- **RAG quality: 80% R@1, 90% R@5, 0.85 MRR.** Built the RAG pipeline (Mistral OCR and embeddings, OpenAI fallback) with an LLM reranker that picks the best client references and explains each choice; 100-question internal eval.
- **Evaluation-driven quality.** Built the platform's evaluation harness (retrieval metrics, LLM-reranker checks) and championed eval-driven development across A^3, surfacing the need for AI governance to the Chief AI Office.

AI Engineer Feb. 2026 – June 2026
Independent — R&D & Open Source Paris, France

- **Open source: codibox on PyPI.** Docker-sandboxed Python execution engine with namespace isolation, a drop-in tool for safe AI code execution in LangGraph.
- **Multi-agent automation: research 3 h → 15 min.** Designed a real estate platform (LangChain deepagents, 6 subagents, 15+ custom tools) for valuation, ROI, and market research.
- **GraphRAG for financial due diligence.** Built a Neo4j platform with hybrid vector/full-text retrieval over a large financial document corpus for LLM due diligence, compliance, and risk Q&A.

AI Engineer · (Fixed-term contract) Feb. 2025 – Feb. 2026
Impactera — Financial Consulting (ESG & data intelligence) Paris, France (Remote)

- **Workflow automation: reports in 30 min, not half a day.** Architected a supervisor-based multi-agent platform (LangGraph, 6 subagents, FastAPI, Docker) spanning document RAG, live API/NL-to-SQL access, web research, and report generation.
- **Reliable RAG: audit-ready answers.** Built a self-correcting pipeline (Qdrant) with semantic chunking, relevance grading, hallucination checks, source attribution, and auto-fallback; shaped by CEO and finance-consultant needs.
- **Multi-LLM routing: 5 providers, auto-fallback.** Engineered routing across OpenAI, Anthropic, Gemini, Groq, and Bedrock with async checkpointing.
- **API orchestration: agents built at runtime.** Built a nested agent that constructs tool-calling agents from OpenAPI specs (Tagetik CPM, client systems), OAuth-secured.

Founding AI Engineer · (Fixed-term contract) June 2023 – June 2024
Tetricks — AI Solutions for Hospitality (École Polytechnique-backed) Remote

- **Inference optimization: 3× throughput, 47% lower cost.** Engineered a vLLM serving stack with AWQ/GPTQ quantization and dynamic batching vs. baseline serving.
- **Vision-LLM delivery: deployed on GCP.** Led integration for automated hotel room descriptions, with the CEO and Dev team; secured Bpifrance innovation funding.
- **Domain fine-tuning: 12+ intents in production.** Fine-tuned Llama 2 on hospitality data for guest intent detection, deployed into production guest-messaging workflows.
- **Multimodal RAG on GCP.** Built a chatbot (Vertex AI + Cloud Run) with a Pydantic-enforced ingestion pipeline and hybrid vector/lexical retrieval with reranking.

ML Research Assistant Jan. 2022 – June 2023
Ibn Tofaïl University — Research Lab Kenitra, Morocco

- **Medical RAG: source-cited clinical Q&A.** Developed a chatbot (Meditron 7B, Qdrant, PubMedBERT) reducing hallucination risk via source-cited retrieval.
- **Dialect classification: 94.5% accuracy.** Built North, a North African dialect classifier (scikit-learn + fine-tuned BERT) for Arabic and Arabizi.
- **Computer vision: 90% average precision.** Benchmarked state-of-the-art architectures on FER2013; 20% generalization gain via augmentation and regularization.

Leadership & Community

Founder & President Oct. 2019 – Jan. 2022
1337 AI — Educational Organisation Ben Guerir, Morocco

- **Community building: 300 members.** Founded and scaled an AI engineering community; ran 2 bootcamps and 6 workshops on Python, ML, and generative AI.
- **Hackathon mentoring.** Co-mentor at ThinkAI 2023 and HackAI 2024–25, alongside Hugging Face and Mistral mentors.

Technical Skills

Agentic AI
LangGraph · LangChain · CrewAI · Google ADK · DSPy · MCP · Agentic RAG

RAG & Document Intelligence
RAG · GraphRAG · Hybrid Search · Reranking · Qdrant · Neo4j · pgvector · Document AI / OCR · VLMs · Multimodal Retrieval

LLM Engineering
vLLM · AWQ/GPTQ · LoRA/QLoRA · Inference Optimization · Multi-LLM Routing · Structured Generation · Fine-Tuning

Evaluation & Observability
RAGAS · LangSmith · Arize Phoenix · LLM-as-Judge · Agent Evaluation · OpenTelemetry · Prometheus · Grafana

AI Systems, Cloud & MLOps
Azure · GCP · AWS · Docker · Kubernetes · Terraform · CI/CD · Azure DevOps · MLflow · GitOps

Security & Governance
LLM Red-Teaming · Guardrails · Prompt-Injection Defense

Programming
Python · FastAPI · Pydantic · SQL · C++ · CUDA · PyTorch

Education

IT Architecture Expert — Data Architecture
RNCP Level 7 · École 42 Paris, Paris, France

Master's in Artificial Intelligence
Ibn Tofaïl University, Kénitra, Morocco

BSc (Licence) in Computer Science
Ibn Tofaïl University, Kénitra, Morocco

Digital Tech Architect — Common Core
École 1337, UM6P, Ben Guerir, Morocco

Baccalauréat, Mathematical Sciences A, Morocco

Academic Experience

Systems & Hardware-Accelerated AI · École 42 Paris Oct. 2024 – Feb. 2025

- Built a **CUDA**-accelerated C/C++ linear algebra engine (matrix ops, inverses, determinants).
- **Inception of Things**: **K3s**/K3D clusters with **ArgoCD** GitOps and GitLab CI/CD.

Certifications

GCP AI Infrastructure Google Cloud 2026
GPU Programming Johns Hopkins 2026
GenAI Winter School École Polytechnique 2024
Red Teaming LLM Apps DeepLearning.AI 2025
LangGraph LangChain 2025
Advanced RAG DeepLearning.AI 2024
Deep Learning Specialization DL.AI 2020

Open Source & Projects

codibox — sandboxed Python exec for agents • PyPI
Real Estate Deep Agents — multi-agent
AgenticDataAnalyst — NL-to-Python

Languages

Arabic (native) · **English** (full professional)
French (professional working)